图书情报工作
第63卷 第14期 2019年7月



# 本地化电子资源使用统计系统解析HTTPS访问数据方法的对比研究

■ 陈广
中国科学院福建物质结构研究所 福州 350002

**摘要：**[**目的/意义**]针对本地化电子资源使用统计系统面临的新问题，提出用于解析HTTPS访问数据的方法并对其进行分析和评价，为图书馆在本地化电子资源使用统计系统中解决基于HTTPS协议访问的电子资源访问数据的采集问题提供参考。[**方法/过程**]从软硬件需求、网络条件、系统功能、用户配合需求4个方面对浏览器扩展程序、支持中间人技术的代理程序和支持SSL代理的网关型设备三种方法进行比较和评价。[**结果/结论**]研究表明，支持中间人技术的代理程序成本适中，系统功能最强，最适宜应用于本地化电子资源使用统计系统。在解决HTTPS访问数据采集问题的基础上，如何保障用户的隐私和数据安全，取得用户的合作和配合是本地化电子资源使用统计系统应用的最大难点。

**关键词：**电子资源 使用统计 浏览器扩展 代理程序 SSL代理网关
**分类号：**G250.7
**DOI**：10.13266/j.issn.0252-3116.2019.14.005



## 1 引言

电子资源是当前图书馆的馆藏重点，电子资源的内容不断扩展、其订购费也逐年攀升。在逐年上涨的经费压力下，图书馆不得不考虑对电子资源进行有效评价，从而为电子资源的订购决策提供参考，最终实现有限经费的最大合理化应用。在电子资源的评价体系中，对其使用情况的统计数据是十分重要的一个评价指标。电子资源使用统计数据对于图书馆具有重要价值，其精确地反映了电子资源的利用状况，可为图书馆重构网络门户、提供用户培训课程以及明确重点突出哪些电子资源产品提供重要参考，还能辅助图书馆员制定有关电子资源购买和管理方面的馆藏决策[1]。

当前图书馆主要通过数据库商提供的基于Counter规范的使用统计报告获取电子资源的使用统计数据，并通过应用ScholaryStats、ExLibrisUStat等支持SUSHI协议的独立统计软件或电子资源管理系统[2]，实现数据收集和统计工作的自动化。

数据库商提供的基于Counter规范的使用统计报告虽然已经十分完善和便捷，但还存在一定的局限性：①数据库商的统计数据有时无法反映真实的用户行为。用户的误操作、重复刷新等行为导致统计用量与真实用量有时并不一致[3]。使用率数据过低时，数据库商亦有可能不提供真实的统计数字[4]。②数据库商的使用统计报告并非实时生成，无法满足图书馆实时查询的需求。③数据库商提供的使用统计报告无法反应图书馆馆藏系统、机构知识库和科学数据库等自建电子资源的使用情况。④Counter报告只能提供统计数字，无法满足图书馆对电子资源进行内容级/用户级的及时、深入的分析和数据挖掘的需要[3]。

为了解决Counter报告的局限性问题，满足图书馆对电子资源进行深入分析和数据挖掘的需求，国内图书馆陆续开展了本地化电子资源使用统计系统的研究工作并取得了丰厚的研究成果。然而近年来本地化的电子资源系统面临一个新的情况，越来越多的数据库为了保障数据安全将电子资源的访问方式从基于HTTP（Hypertext Transfer Protocol）协议转换为基于HTTPS（Hyper Text Transfer Protocol over Secure Socket Layer）协议，如OSA数据库于2017年，ACS和Wiley数据库于2018年分别部署了基于HTTPS协议的访问并取消了基于HTTP的访问；ScienceDirect、Nature、Springer等数据库则早在2017年之前就采用了基于HTTPS协议

作者简介：陈广（0000-0002-3828-7401），馆员，E-mail：chenguang@fjirsm.ac.cn。
收稿日期：2018-11-08 修回日期：2019-01-15 本文起止页码：36-43 本文责任编辑：杜杏叶

的访问。HTTP 协议采用明文传输数据,HTTPS 协议采用密文传输数据,这就使得原先适配于 HTTP 协议的本地化电子资源使用统计系统无法采集更换协议后的电子资源的访问数据。如何采集基于 HTTPS 协议访问的电子资源的访问数据,成为本地化电子资源统计系统急需解决的重要问题。

针对上述问题,本文提出三种可以用于解析 HTTPS 访问数据的方法,综合对比分析这三种方法在软硬件需求、网络条件、系统能力和用户配合需求四个方面的优劣势,为图书馆解决基于 HTTPS 协议的电子资源访问数据的采集问题提供参考。

## 2 背景分析

### 2.1 本地化电子资源使用统计系统研究现状

国内图书馆很早就开展了本地化电子资源使用统计系统的研究。从采用的技术方法来看,本地化电子资源统计系统主要可以分为两种:基于网关日志的采集分析模式和基于旁路监听的采集分析模式。在网关日志的采集分析方面主要有:使用网关日志构建电子资源使用统计系统[6];挖掘防火墙日志构建电子期刊数据库统计分析系统[7];通过代理服务器的 Web 日志构建电子资源日志统计系统[8];图书馆数字资源访问系统的日志处理和数据挖掘[9]等。在旁路监听的采集分析模式方面主要有:基于 ERU 系统研究图书馆用户信息行为数据采集方法[10],利用旁路监听设计及应用电子资源访问管理与控制系统[11],利用旁路监听设计和应用高校电子资源访问管理控制系统[12],基于旁路监听设计和实现数字资源评估系统[13]等。

在基于网关日志的采集分析模式下,网关设备如核心交换机,防火墙,代理服务器等,会对流经网关的互联网访问数据进行记录并形成日志文件,日志文件中包含了用户对电子资源的访问数据。在这种情况下,图书馆只要采取一定的日志收割策略,通过编写日志收割程序、过滤和分析日志信息之后就能生成电子资源的使用统计报告。基于网关日志的采集分析模式的优点在于不需要变更现有的网络拓扑,也不需要增加硬件设备,可以直接通过网关设备自带的日志功能来获取电子资源的使用统计报告。局限性则在于网关日志记录的信息可能不够完整,无法满足图书馆深入分析和挖掘数据的需求。使用统计报告的生成频率,取决于日志文件的生成频率以及相应日志收割程序的转换性能,难以做到使用统计报告的实时生成,也无法对电子资源的使用情况进行实时监控。由于只是采集了日志信息,在用户发生违规的电子资源使用行为时,并不能及时终止违规行为。

基于旁路监听的采集分析模式是在现有网络拓扑的基础上增加专门的数据分析服务器,并将数据分析服务器与网络出口的网关设备相连接,在网关设备上通过端口镜像功能将数据包复制至数据分析服务器。数据分析服务器捕获并解析数据包,对数据包内容进行过滤分析后生成电子资源使用统计报告。这种模式的优点在于使用端口镜像功能复制了数据包,不需要改变原先的网络拓扑结构,也不会对用户的访问行为造成影响;可以实时监控电子资源的使用情况从而对用户的违规信息行为进行预警;获得的电子资源使用数据完整准确,能够满足图书馆深入挖掘数据的需求。局限性则在于需要增加专门的数据分析服务器用于监听、采集分析数据,成本较高;尽管可以对违规行为进行实时监控,但旁路监听的方式并没有参与用户的访问行为,违规行为发生时同样无法立即终止用户的违规行为。

### 2.2 HTTP 与 HTTPS

HTTP 超文本传输协议是在互联网上进行通讯时使用的协议方案。HTTP 是无状态、简单快速、基于 TCP 的可靠传输协议,其最主要的应用是 Web 浏览器和 Web 服务器之间的双工通信[14]。目前互联网上的 Web 服务器基本都使用 HTTP 协议来传输数据,大部分电子资源的访问也都基于 HTTP 协议。

HTTP 虽然方便快捷,却存在数据安全方面的问题。HTTP 采用明文传输数据,其传输的数据对整个传输链路上的网络设备都是透明的,这就使得第三方可以窃听或篡改数据,甚至可以冒充 Web 服务器的身份来同用户进行通信。

为了解决 HTTP 协议的数据安全问题,NetScape 公司设计了 SSL(Secure Sockets Layer)协议用于对 HTTP 协议传输的数据进行加密并将 SSL 应用在了自家的浏览器中,从而诞生了 HTTPS。SSL 协议总共有 3 个版本,目前最新的为 SSL3.0 版本。1999 年互联网标准化组织 ISOC 接替 NetScape 公司,发布了 SSL 的升级版 TLS 1.0 版。TLS 经历了两次升级,目前最新的是 2011 年发布的 TLS1.2 的修订版。SSL 及其继任者 TLS 是为网络通信提供安全及数据完整性的一种安全协议,其主要的作用是:认证用户和服务器,确保数据发送到正确的客户机和服务器;加密数据以防止数据中途被窃取;维护数据的完整性,确保数据在传输过程中不被改变[15]。

从图1来看，HTTP和HTTPS最大的区别是HTTPS在HTTP的基础上引入了安全层用于加密数据，安全层可选用SSL或TLS协议。数据在到达传输层的时候就已经加密完成，整个数据链路上传输的数据都是加密的，避免了数据被修改和篡改的可能。HTTPS增加了安全层使得其部署的成本相比于HTTP更高，所占用的服务器资源更多，耗费的访问时间更长。尽管部署HTTPS要求的条件较高，更多的电子资源提供商为了自身数据的安全逐渐开始采用基于HTTPS协议的访问取代原先的基于HTTP协议的访问。




图1 HTTP与HTTPS网络模型

## 3 本地化电子资源使用统计系统面临的问题及解决方案

对于采用HTTPS协议访问的电子资源，无论是基于网关日志的模式还是基于旁路监听的模式，当访问数据到达网关或数据分析服务器时都已经是加密后的密文，这就使得网关或数据分析服务器无法获得数据的详细内容，只能获取到用户IP地址、服务器IP地址和域名三个信息，这些信息并不足以生成电子资源的使用统计报告。现有的本地化电子资源使用统计系统不再适用于新的电子资源访问方式，急需解决基于HTTPS协议访问的电子资源使用数据的采集问题。

从HTTPS的工作原理来看，想要解密HTTPS可以采用2种方式：①HTTPS的加密和解密工作是在安全层进行的，对应用层而言数据还未加密或已经解密完成，只要在应用层布置监听程序就可以得到未加密的数据。②使用中间人(Man-in-the-middle)技术来控制客户端和服务器端之间的数据通讯。在客户端和服务器端中间添加第三方，第三方分别与服务端和客户端建立连接，自身扮演客户端同真实服务端通讯，同时扮演服务端同真实客户端通讯，交换客户端和服务端的数据，使得通讯两端都认为自己直接与对方对话，事实上整个会话都被第三方所控制[16]。

### 3.1 浏览器扩展程序

对于第一种方式，浏览器是访问Web资源最主要的工具，工作于应用层。所有的网络访问数据对浏览器而言都是透明的，可以直接以明文方式查看，通过对浏览器的通讯数据进行监听就可以获取基于HTTPS协议访问的数据。这种方法事实上是绕过了安全层，在应用层上监听数据。

主流浏览器都支持扩展(Extension)程序，扩展程序是用来修改Web浏览器功能的代码，使用标准的Web技术(JavaScript、HTML和CSS)和一些专用JavaScript APIs编写，能够实现网络请求控制，各类事件监听等功能。目前主流的浏览器名称、对应的内核信息以及支持的扩展接口如表1所示：

表1 主流浏览器名称、内核和扩展支持

| 浏览器 | 内核 | 扩展接口 |
|---|---|---|
| IE浏览器 | Trident内核 | BHO(Browser Helper Object) |
| Chrome浏览器、Opera浏览器 | Blink内核 | Chrome Extension |
| Microsoft Edge浏览器、Safari浏览器 | Webkit内核 | Webkit Extension |
| Firefox Quantum浏览器 | Quantum内核 | WebExtensions API |
| 360浏览器、猎豹浏览器、 | Trident + Blink双内核 | BHO + Chrome Extension |
| 腾讯TT、淘宝浏览器、搜狗浏览器、傲游浏览器、百度浏览器、世界之窗浏览器 | Trident + Webkit双内核 | BHO + Webkit Extension |

表1中的浏览器使用了Trident、Webkit、Blink和Quantum 4种内核，采用相同内核的浏览器其扩展程序可以相互兼容，Blink内核为Webkit内核的升级版本，这两种内核的扩展程序也可以相互兼容。Firefox Quantum浏览器的WebExtensions API可以兼容Chrome Extension，因此在制作浏览器扩展程序时仅需针对BHO和Chrome Extension编写即可兼容市面上主流的浏览器。以WebExtensions API为例，从Firefox Quantum浏览器官方提供的文档可以得知WebExtensions API中提供WebRequest模块，该模块中的onBeforeRequest方法在浏览器发送请求时触发，onCompleted方法在浏览器请求完成时触发[17]。通过在这两个方法中加入监控代码，制作相应的扩展程序，并将扩展程序安装在用户的浏览器中，在用户访问电子资源时，将用户发送的请求信息和服务器返回的内容同步发送一份到图书馆的服务器，就可以实现基于HTTPS协议访问的

陈广.本地化电子资源使用统计系统解析 HTTPS 访问数据方法的对比研究[J].图书情报工作,2019,63(14):36-43.

电子资源访问数据的采集。

## 3.2 中间人技术

使用中间人技术解密 HTTPS 协议有两种方法。第一种是使用支持 SSL 代理的网关型设备替代原来的网关,或将其添加到原来的网络拓扑中。支持 SSL 代理的网关设备利用 SSL 代理证书替换加密 Web 网站的数字证书,并将 SSL 代理证书发送到客户端的 Web 浏览器,在此过程中,设备分别作为 SSL 客户端和 SSL 服务器与 Web 服务器和 Web 浏览器建立 SSL 连接,从而获得加密通信的明文内容。SSL 代理证书是使用设备本身的证书对 Web 服务器证书重新签发而成的证书[18]。整个工作过程如图 2 所示:

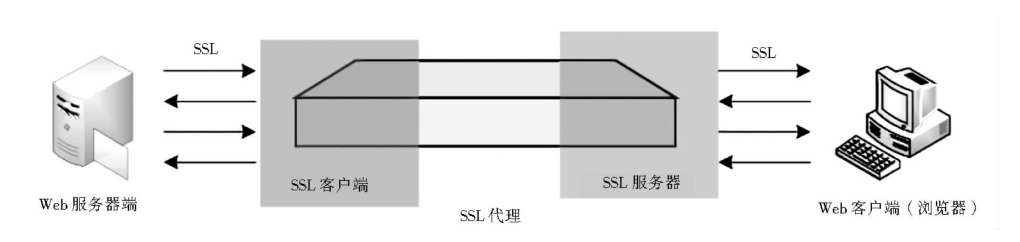



图 2　SSL 代理网关

这种方法事实上还是基于网关日志的采集分析模式,只是将原先不支持 HTTPS 解析的网关设备替换为支持 HTTPS 解析的网关设备,仍然需要编写日志收割程序来生成电子资源使用统计报告。由于审计方面的需求,越来越多的网关设备开始支持 SSL 代理功能。

另外一种方法是在数据分析服务器上部署支持中间人技术的代理程序如 Fiddler、Charles、whistle 等,替代原有的数据包捕获程序,用于抓取和分析数据包。以 Charles 代理程序为例,其解析 HTTPS 协议的过程如图 3 所示:

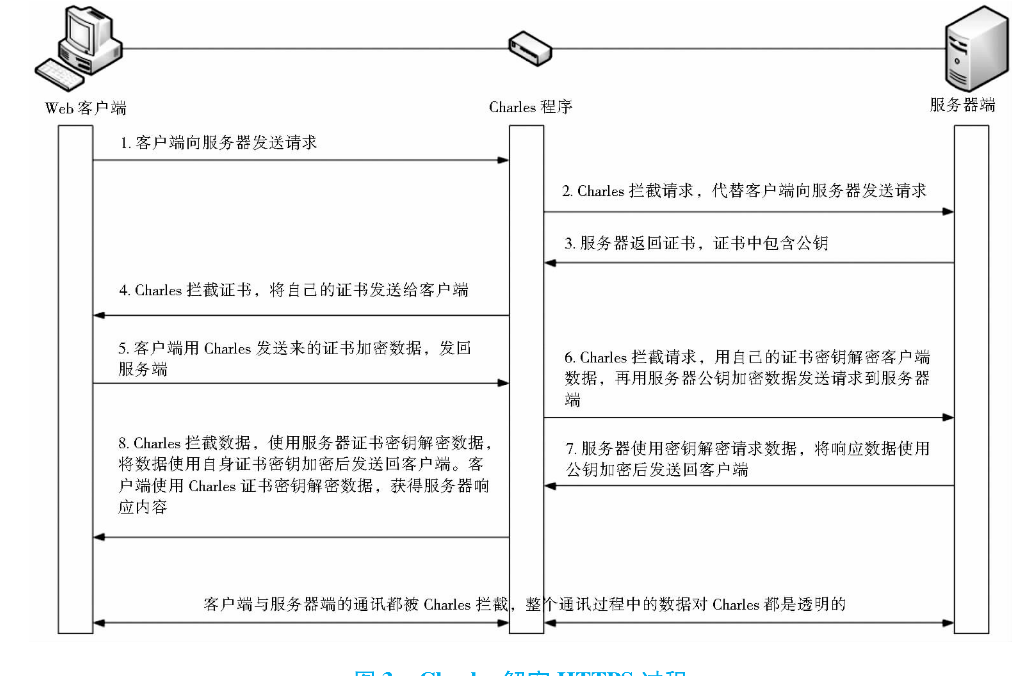



图 3　Charles 解密 HTTPS 过程

从图 3 可以得知 Charles 代理程序在整个通讯过程中掌握了服务器证书公钥和 HTTPS 连接的对称密钥,所有的密文都可以使用对应密钥来进行解密,整个通讯过程对 Charles 是透明的。通过在支持中间人技术的代理程序中添加监听代码,就可以实现基于 HTTPS 协议访问的电子资源访问数据的采集。

## 3.3 数据分流方法

支持 SSL 的网关设备和代理程序都是使用中间人技术来解析 HTTPS,中间人技术需要能够直接同客户端和服务端进行通讯,这就要求将采用中间人技术的设备或代理程序接入到原有的网络拓扑中。对于部署代理程序的数据分析服务器而言,如果将其直接接入网络拓扑,除了需要解析、转发电子资源的访问数据之外还需要负载非电子资源的访问流量,数据分析服务器并非专门的网关设备,有可能因为负载太大而导致服务器无法正常工作。为了解决这个问题,可以采用

数据分流的方法只将电子资源的访问数据转发至数据分析服务器，而非电子资源的访问数据则直接将发送至出口网关。

最常用的数据包转发技术为策略路由（Policy-Based Routing，PRB），策略路由是一种依据用户制定的策略进行路由选择的机制。通过在核心交换机上开启策略路由功能，创建路由访问控制列表（Access Control List，ACL），将电子资源服务器的 IP 地址信息存储在 ACL 列表中，配置好转发策略后就可以实现电子资源访问数据的分流。当网络访问数据到达核心交换机后，将访问数据包的目标 IP 地址与 ACL 列表进行匹配，目标 IP 地址为电子资源服务器 IP 地址的将数据包转发至数据分析服务器，其它访问数据则直接发送至出口网关。策略路由的好处是数据包转发是在核心交换机上进行的，不需要用户参与，用户无法感知到策略路由带来的变化。

策略路由也存在一定的隐患，策略路由转发数据的依据是目标服务器的 IP 地址。部分电子资源为了加速自身的访问采用 CDN（Content Delivery Network，内容分发网络）技术，通过布置多台缓存服务器，将这些缓存服务器分布到用户访问相对集中的地区或网络中，在用户访问网站时，利用全局负载技术将用户的访问指向距离最近的工作正常的缓存服务器上，由缓存服务器直接响应用户请求[19]。采用 CDN 技术的电子资源配置有多个缓存服务器，相应就有多个 IP 地址，当所有电子资源的 IP 地址数量超过 ACL 列表所能支持的最高数量时，就会导致部分电子资源的访问数据无法被转发到数据分析服务器，从而导致电子资源使用统计数据的不准确。

另外一种数据分流的方法是 PAC（proxy auto-config，代理自动配置），PAC 是一个自动代理配置脚本文件，它能够决定浏览器访问网络资源时走默认通道还是代理服务器通道。浏览器通过配置 PAC 文件来实现自动代理功能。PAC 文件中包含一个 JavaScript 形式的函数 FindProxyForURL，该函数可返回包含一个或者多个访问规则的字符串，这些规则字符串决定浏览器是否通过代理程序访问网络资源。在 FindProxyForURL 函数中对浏览器访问的 URL 进行判断，当 URL 包含电子资源的域名信息时，指定其通过代理程序访问，从而实现电子资源访问流量的分流。相比于策略路由，PAC 使用域名信息来分流数据，所需要的匹配规则大幅减少，并且不需要在网关设备上做任何设置，更适合作为数据分流的方法，但 PAC 需要用户在浏览器中配置 PAC 文件，有可能造成用户的抵触。

## 4 三种解析 HTTPS 方法的对比分析

上述的三种方法（以下分别简称为：浏览器扩展、SSL 网关和代理程序）都可以用于解析基于 HTTPS 协议的访问数据，但每种方法所适用的环境不尽相同。本文将从软硬件需求、网络条件、系统功能、用户配合需求四个方面对这三种方法进行比较，分析不同方法的优劣势。

### 4.1 软硬件需求

软硬件需求是指应用各种方法时需要添加的硬件设备，安装的软件以及需要自行编写的程序和文件。

在硬件方面三种方法都需要一台专门的服务器用于保存电子资源的访问记录并生成使用统计报告。SSL 网关还需要购买专门的网关设备，所需成本最高。代理程序需要分析和转发数据包，对服务器的性能要求较高，硬件成本取决所需处理的流量大小，但一般不会超过 SSL 网关。浏览器扩展仅需在服务器上架设 Web 服务器用于接收浏览器采集到的访问信息，对服务器的性能要求最小，所需的硬件成本最低。

在软件方面三种方法都需要安装数据库程序来存储用户请求信息。SSL 网关需要编写日志收割程序；浏览器扩展需要编写扩展程序，同时编写 Web 页面用于接收访问请求信息；代理程序需要安装相应的代理软件，并在软件中编写代码用于记录电子资源访问请求；采用数据分流技术的代理程序还需要配置核心交换机的策略路由功能或制作 PAC 文件。从技术难度上来说，代理程序所需的技术难度最高，浏览器扩展其次，SSL 网关最低。

### 4.2 网络条件

网络条件是指方法的应用是否需要变更原有的网络拓扑，服务器或网关设备需要安装在网络当中的哪个位置。

SSL 网关对网络环境的要求最高，需要将 SSL 网关直接接入到原有的网络拓扑中，作为网络拓扑的核心节点。不采用数据分流方法的代理程序所在的数据分析服务器和 SSL 网关一样需要接入到网络拓扑中充当核心节点。采用策略路由方法分流数据的代理程序需要将数据分析服务器连接在核心交换机上，数据分析服务器要能够与核心交换机直接通讯，中间不能有任何额外的网络节点。采用 PAC 分流的代理程序理论上来说可以将服务器布置在单位内部网络的任意位置，只要保障用户访问数据可以被转发至服务器。一

般为了保障访问速度,减少中间节点,还是应该将服务器与核心交换机直接连接。浏览器扩展对网络环境的要求最低,服务器能够接收到用户发送来的信息即可,甚至可以将服务器架设在外部网络。

### 4.3 系统功能

系统功能可以从数据采集的完整性,使用统计报告生成的及时性,系统控制能力三个方面进行评价。

数据采集的完整性方面,SSL 网关的数据来源于网关日志,数据采集的完整性取决于网关日志的信息量,通常情况下网关日志只记录 URL、来源 IP、访问时间几个信息,记录的信息量较少,数据采集的完整性较差。代理程序和浏览器扩展都可以直接获取到整个访问行为的全部数据,包括浏览器类型、用户 IP 地址、目标 IP 地址、访问时间、访问页面 URL、访问页面 HTML 内容、下载资源类型等,甚至可以通过用户 IP 地址再获取到用户姓名和所在部门等信息,采集到的数据信息十分完整。

使用统计报告生成的及时性方面,SSL 网关的使用统计报告的生成频率取决于日志收割程序的采集频率,通常的做法是按天或按小时收割日志,及时性方面为小时级别。代理程序和浏览器扩展记录访问数据是随着用户的信息访问行为实时发生的,也就是说可以实时监控用户的访问行为以及实时生成使用统计报告,及时性为实时级别。

系统控制能力方面,SSL 网关一般只对访问行为进行记录,不提供更改数据内容的能力,有些 SSL 网关提供黑名单功能,可以拦截指定域名的访问,系统控制能力一般。代理程序和浏览器扩展对用户的访问行为具有完全的控制能力,可以拦截或修改用户请求和服务器的返回内容,系统控制能力十分强大。浏览器扩展需要安装在用户的浏览器上,其实现的功能要考虑普适性问题,难以针对具体用户做特定的功能控制,当功能出现变更的时候要等待用户更新浏览器扩展后新的功能才能生效。代理程序对用户请求的控制在数据分析服务器上进行,可以针对不同情况设定更多的功能,也可以针对特定用户请求设定不同的应对策略,当系统功能更新后可以立即生效,因此代理程序的系统控制能力要高于浏览器扩展。

### 4.4 用户配合需求

用户配合需求是指需要用户在客户端做出的配合,比如安装证书、安装浏览器扩展程序、修改系统设置等。

采用中间人技术的 SSL 网关和代理程序需要将 SSL 代理证书发送给用户浏览器用于加密数据,通常这个证书都是由 SSL 网关或代理程序自行签发。浏览器接收到证书后会对证书做验证,当浏览器发现证书并非由受信任的根证书机构签发时,会在浏览器访问相关域名时提示安全证书不受信任,对应的电子资源也会出现无法正常访问的情况。为了解决这个问题,就需要用户在客户端浏览器中导入 SSL 代理证书的根证书,并将该证书添加到受信任的根证书颁发机构中。由于用户已经手动信任 SSL 代理证书的根证书,相应的 SSL 代理证书也会被浏览器所信任,电子资源就可以正常访问。采用 PAC 分流方法的代理程序还需要用户在本地浏览器中配置 PAC 的相关设置,使自动代理功能可以生效。浏览器扩展则需要用户将扩展程序安装到浏览器中,并开启扩展程序功能,从而使扩展程序可以正常工作。三种方法都需要用户做出配合才能实现对基于 HTTPS 协议访问的电子资源数据的采集。

综上所述三种方法的对比分析如表 2 所示:

表 2　三种方法对比分析

| 项目 | SSL 网关 | 代理程序 | 浏览器扩展 |
|---|---|---|---|
| 软件技术难度 | 低 | 高 | 中 |
| 硬件需求成本 | 高 | 中 | 低 |
| 网络环境要求 | 高 | 中 | 低 |
| 系统功能 | 低 | 高 | 中 |
| 用户配合需求 | 中 | 中 | 高 |

## 5 结语

### 5.1 三种方法的选取策略

SSL 网关需要添加或更换网关设备,改变原来的网络拓扑结构,硬件成本最高,系统功能最低。如果图书馆对电子资源的使用统计报告要求不高,并且所在机构正好需要对网关设备进行升级换代时,可以建议网络部门选取支持 SSL 代理的网关设备。

浏览器扩展所需的成本最低,其提供的系统功能已经可以满足大部分图书馆的需求。浏览器扩展需要在客户端浏览器中安装扩展程序,部分用户可能因为安全问题而拒绝安装。当图书馆对用户有较强的控制能力,能够强制用户安装浏览器扩展,在成本有限的情况下,可以采用浏览器扩展的方法。

代理程序所能提供的系统功能最强,但技术方面的要求最高,同时也需要网络部门和用户给予一定的配合。如果图书馆需要十分强大的系统功能,要求系统能够针对不同用户采取不同的采集和控制策略,系统生成的使用统计报告详细准确,并且图书馆自身具

有一定的技术开发能力,能够取得用户的配合,可以采用代理程序的方法。

以笔者所在单位中国科学院福建物质结构研究所(以下简称“我所”)为例,在构建电子资源使用统计系统时考虑了以下因素:①我所规定课题组根据电子资源使用量分担一部分电子资源费用,这就要求电子资源使用统计系统能够提供个人以及课题组的准确统计数据;②电子资源使用统计系统能够监控、预警、处理异常的文献下载行为;③用户能够接受安装 SSL 代理证书,无法接受浏览器扩展程序和 PAC 文件;④我所的核心交换机和防火墙设备支持策略路由功能,网络部门愿意配合配置策略路由功能。综合上述因素,我所选取了 Fiddler 代理程序用于解析记录电子资源访问数据,并在防火墙设备上开启策略路由功能分流电子资源访问数据。

### 5.2 我所电子资源使用统计系统实施效果

根据上述方案,我所于 2017 年 7 月份完成本地化电子资源使用统计系统(以下简称“系统”)部署。经多次修改调整后,目前系统运转稳定并提供了良好的服务。

在数据采集方面,系统实现了 ACS、Wiley、Science Direct、RSC、Nature、Science、AIP、OSA、Springer 共计 9 个数据库的全文访问数据的记录,2018 年度共存储全文访问记录 431 342 条。通过该数据可以生成各数据库的全文下载量、下载量占比和篇均成本数据,为图书馆掌握电子资源使用情况,调整电子资源保障策略提供了强有力的支撑。除此之外,还可以了解各课题组的学科方向、文献需求等信息,有助于图书馆开展个性化信息服务。

在数据应用方面,根据 2017 年 12 月 1 日至 2018 年 11 月 30 日的电子资源全文访问数据,我所于 2018 年 12 月份完成了课题组分担电子资源费用的工作。全所 92 个课题组合计分担电子资源费用1 002 090.27 元。由于系统记录的全文访问数据包含用户上网账号,IP 地址,全文 URL 和访问时间等信息,数据详细准确,各课题组对数据和费用没有异议,电子资源费用分担工作得以顺利进行。

在系统控制方面,用户在 ACS 数据库点击文章标题时会自动跳转到全文页面并生成一次全文访问记录,当用户在全文页面再下载 PDF 文件时会再生成一次全文访问记录,这就造成了全文数据的重复访问。针对这种情况,系统在 Fiddler 程序中采取了 URL 替换的措施。当系统检测到用户点击 ACS 标题时,将全文页面 URL 替换为对应的文摘页面 URL,用户点击 ACS 数据库文章标题时就不再访问全文页面而是访问文摘页面,从而避免了不必要的全文数据的重复访问。

### 5.3 存在的问题及未来发展方向

SSL 网关和代理程序需要用户安装证书,浏览器扩展需要用户安装扩展程序,PAC 分流方法需要用户在浏览器中配置 PAC 文件,无论哪种方法都需要在用户在客户端添加额外的文件,不可避免的侵犯了用户的隐私,同时也给用户带来数据安全方面的风险。用户出于数据安全和个人隐私方面的考虑,一般都难以接受在客户端安装额外的程序。图书馆在应用本地化电子资源使用统计系统的过程中除了应该考虑如何保护用户的隐私不被侵犯,还应保障不会因为安装证书或扩展程序而带来用户数据泄漏的风险。

本地化电子资源使用统计系统在未来应该考虑引进新的技术手段,在不需要用户安装任何程序的情况下实现基于 HTTPS 协议的电子资源访问数据的采集。系统采集的到的数据不应该只用于生成使用统计报告,不应该仅是采用可视化的方法展示数据,而应采用数据分析,机器学习,深度学习等方法对数据进行深入挖掘,分析了解用户需求,构建用户画像,为个性化的知识服务提供依据。

**参考文献:**

[1] 陈大庆,叶兰,杨巍,等. 电子资源使用统计平台 USSER 的设计与实现[J]. 图书情报工作,2015,59(1):106-112.

[2] ANDERSON E K. Electronic resource management systems and related products[J]. Library technology reports,2014, 50(3):30-42.

[3] 王丹丹. 数字图书馆用户使用数据统计的现状与趋势研究[J]. 图书馆建设,2012(11):66-69.

[4] TRIPATHI M,JEEVAN V. A selective review of research on e-resource usage in academic libraries[J]. Library review,2013,62(3):134-156.

[5] 朱玲,崔海媛. 高校图书馆电子资源使用监控与统计系统数据获取质量评估方法探讨[J]. 图书情报工作,2016,60(5):51-57.

[6] 闫晓弟,邵晶,周奇,等. 电子资源利用统计网关系统的设计与实现[J]. 现代图书情报技术,2008,24(8):97-100.

[7] 王孝亮,王威. 通过防火墙日志挖掘构建电子期刊数据库统计分析系统[J]. 现代图书情报技术,2013,29(S1):122-126.

[8] 郭振英,赵文兵,魏育辉. 电子资源日志统计系统分析与设计[J]. 现代图书情报技术,2008,24(9):102-106.

[9] 周欣,陆康. 基于图书馆数字资源访问系统的读者行为数据挖掘研究[J]. 现代情报,2016,36(1):51-56.

[10] 张计龙,殷沈琴,陈铁. 基于 ERU 的图书馆用户信息行为数据采集方法研究——以复旦大学图书馆为例[J]. 图书馆杂志,

2014,33(12):10 - 16.
[11] 邹荣,张成昱,姜爱蓉,等. 电子资源访问管理与控制系统的设计及应用[J]. 图书情报工作,2010,54(1):121 - 124.
[12] 施晓华,钱吟,谢锐. 高校电子资源访问管理控制系统的设计和应用[J]. 计算机应用研究,2011,28(3):1042 - 1045.
[13] 王政军,董晓梅,俞小怡. 基于旁路监听的数字资源评估系统的设计与实现[J]. 图书情报工作,2015,59(9):52 - 57.
[14] GOURLEY D, TOTTY B. HTTP 权威指南[M]. 陈涓,赵振平,译. 北京:人民邮电出版社,2012.
[15] YANG A. SSL 详解[EB/OL]. [2018 - 11 - 07]. https://www.cnblogs.com/NathanYang/p/9183300.html.
[16] QU J. 三种解密 HTTPS 流量的方法介绍[EB/OL]. [2018 - 11 - 07]. https://imququ.com/post/how-to-decrypt-https.html.
[17] fkyq01. webRequest [EB/OL]. [2018 - 11 - 07]. https://developer.mozilla.org/zh - CN/docs/Mozilla/Add - ons/WebExtensions/API/webRequest.
[18] SSL 代理 [EB/OL]. [2018 - 11 - 07]. http://docs.hillstonenet.com/cn/Content/9_Security/SSLProxy.htm.
[19] 123freya321. cdn 之高速缓存服务器的搭建和配置 [EB/OL]. [2018 - 11 - 07]. https://blog.csdn.net/zxy15771771622/article/details/79310601.

## A Comparative Study of Local Electronic Resource Usage Statistics System for Resolving HTTPS Access Data

Chen Guang
Fujian Institute of Research on the Structure of Matter, Chinese Academy of Sciences, Fuzhou 350002

**Abstract:** [**Purpose/significance**] Aiming at the new problems of local electronic resource usage statistics system, this paper proposes some methods for resolving HTTPS access data, analyzes and evaluates these methods. It provides a reference for the library to solve the collection problem of electronic resource access data based on Https protocol access in the local electronic resource usage statistics system. [**Method/process**] From the four aspects of hardware and software requirements, network condition, system functions and user coordination requirements, this paper compared and evaluated the three methods including browser extensions, proxy supporting MITM technology, and devices that support SSL proxy gateway. [**Result/conclusion**] The result shows that proxy supporting MITM technology is moderately costly and has the strongest system functions, which most suitable for local electronic resource usage statistics systems. On the basis of solving the problem of HTTPS access data collection, how to ensure user privacy and data security, obtain user cooperation and cooperation will be the biggest difficulty in the application of local electronic resource usage statistics system.

**Keywords:** electronic resource usage statistics browser extension proxy SSL proxy gateway



## 下 期 要 目